Journal of Data Science 0 (0), 1–21
??? 2026
DOI: 10.6339/26-JDS1239
Data Science in Action

Predictive Modeling of Short-Term Recidivism: A Mixture Cure Rate Approach with Diverse Link Functions

Chaeyeon Yoo[1,*] and Dipak K. Dey[1]
[1]*University of Connecticut, Storrs, USA*

Abstract

Addressing short-term recidivism presents a significant challenge within the criminal justice system, necessitating robust predictive models that can accurately identify risk factors and predict outcomes. This paper introduces application of Bayesian cure rate models to predict short-term recidivism using data from individuals released from Iowa prisons in 2018. The mixture cure rate models effectively capture the recidivism risk by distinguishing between cured individuals, those unlikely to re-offend, and those susceptible to relapse. By implementing accelerated failure time models combined with logistic regression and various link functions, including logit, skewed logit, reversed power logit, and flexible generalized logit, we enhance the model's flexibility and interpretability. Through a comparative analysis of these models, key predictors of recidivism are identified and discussed in detail.

Keywords *accelerated failure time; Bayesian inference; generalized logistic distribution; time-to-event*

1 Introduction

Addressing recidivism, which is defined as re-offending after criminal justice system intervention, presents a significant challenge requiring robust models that can identify risk factors and predict outcomes. The U.S. Bureau of Justice Statistics reports that approximately 68% of released prisoners are rearrested within three years and 83% within nine years (Alper et al., 2018), a trend confirmed by recent 34 states studies showing 62% re-arrest rate within three years (Durose and Antenangeli, 2021). These persistent rates of recidivism impose substantial costs on public safety, driving interest in statistical modeling based decision making. While overall recidivism rates are considerably high, these rate can change depending on the factors. For instance, re-offending rates reach 72% for individuals aged 24 or younger, compared to only 21% for those 65 or older. Furthermore, with public order re-offense rating 54% is nearly double the rate of violent re-offenses. This variation among the factors motivates to develop quantitative methods that can analyze the complex patterns across diverse offender subgroups.

In response to these challenges, modern systems have not only relied on actuarial risk assessment tools such as COMPAS (Dressel and Farid, 2018), which integrate statistical models with administrative data to guide supervision policies, but have also seen the development of a wide range of modern statistical approaches to predict the risk of recidivism. Classification models, ranging from logistic regression to machine learning techniques such as decision trees, have been widely used in criminology to predict recidivism risk and to construct statistical

*Corresponding author. Email: chaeyeon.yoo@uconn.edu.

Received April 9, 2025; Accepted June 27, 2026

risk assessment tools used in criminal justice decision making (Berk, 2016; Tollenaar et al., 2016; Ting et al., 2018). While these tools excel in binary classification, the question of when someone will re-offend has gained recognition, emphasizing a shift toward time-to-event modeling.

Building on the early work of Partanen (1969), survival models have been widely used to study the timing of events in criminological literature. However, standard frameworks implicitly assume that all individuals will eventually recidivate if followed long enough. To address this, criminological research has proposed split-population models, which account for the possibility that a subset of individuals may never re-offend (Schmidt and Witte, 1989, 2012). These are equivalent to mixture cure rate models, which partition the population into a susceptible and an effectively immune group (Maltz, 1984; Maller and Zhou, 1996). Although widely used in medical statistics, these models remain relatively underexplored in criminological research despite their natural fit for modeling long-term desistance. Recent work by De la Cruz et al. (2022) introduced a Bayesian mixture cure rate model tailored to recidivism analysis, enabling simultaneous estimation of short-term and long-term re-offending risks.

Mixture cure rate models consist of two key components: an incidence model, estimating the probability of belonging to the susceptible group, and a latency model that captures the survival distribution among susceptible individuals. Since their introduction in oncology studies by Boag (1949) and Berkson and Gage (1952), these models have evolved from parametric formulations, coupling logistic regression with Weibull or accelerated failure time (AFT) models (Farewell, 1977, 1982; Yamaguchi, 1992), to semiparametric frameworks utilizing Cox proportional hazards and semiparametric AFT structures (Kuk and Chen, 1992; Sy and Taylor, 2000; Li and Taylor, 2002; Zhang and Peng, 2007).

While most mixture cure rate studies focus on the latency component, the cured proportion is typically handled by standard logistic regression. However, defaulting to symmetric logit link functions can be inadequate for real world data with highly unbalanced distribution. In recidivism prediction, the binary incidence component often exhibits substantial class imbalance between recidivists and non-recidivists, which can bias estimation when symmetric links are applied. To address this, several flexible link functions have been proposed, including the scobit model (Nagler, 1994), generalized t-link models (Chen et al., 1999; Kim et al., 2008), power and reversed power links (Bazán et al., 2017), and flexible generalized logit links (Prasetyo et al., 2020). Despite these developments, the use of such flexible link functions within mixture cure rate models has received limited attention.

In this paper, we present a Bayesian mixture cure rate model tailored for short-term recidivism prediction. We extend the AFT mixture cure rate model by integrating flexible link functions for logistic regression: logit, skewed logit, reversed power logit, and flexible generalized logit, which have not been explored in recidivism modeling. These link functions provide improved flexibility and interpretability in estimating the probability of re-offending, in the presence of skewed group membership. Although this study focuses on extensions of the logit family, the framework can readily accommodate alternatives such as the probit or cauchit links, offering further avenues for model generalization.

The data used in this article consist of information on a cohort of release from the Iowa prison system in 2018, with information on demographics, original offense, and risk level. This data set, available from the Iowa government correctional system database (https://data.iowa.gov/Correctional-System/Iowa-Prison-Recidivism-and-Change-by-Cohort/dnzw-paxg/about_data). Bayesian methods are employed for inference and model fitting with `stan`, and model performance is evaluated through Deviance Information Criterion (DIC), Leave-

One-Out Information Criterion (LOOIC) and Watanabe-Akaike Information Criterion (WAIC) (Spiegelhalter et al., 2002; Watanabe, 2010; Vehtari et al., 2017).

The remainder of this paper is organized as follows. Section 2 provides a detailed description of the Iowa 2018 short-term recidivism dataset used in this study. Section 3 introduces the framework of the mixture cure rate model, highlighting both incidence and latency components. Section 4 presents the link functions used in the incidence model, including flexible and asymmetric alternatives to the traditional logit link. Section 5 discusses the model assessment criteria used to evaluate predictive performance. Section 6 presents simulation studies to assess the behavior of the proposed models under controlled conditions. Section 7 applies the methodology to the Iowa recidivism data, providing insight into the practical implications of our approach. Finally, Section 8 concludes with a summary of findings and directions for future research.

2 Data Description

The Iowa 3-year recidivism dataset consists of 5,010 individuals from the 2018 prison-release cohort followed through 2021 (Iowa Department of Corrections, 2024). The data tracks successful community reentry or recidivism, recording the specific time-to-event for those returning to a correctional facility, with a total observed recidivism rate of 38.7% over the three-year follow-up period. The dataset encompasses information on age, sex, race, original offense committed, and the individual's level of risk, all of which will be considered as covariates in the model. Table 1 and Table 2 provides the numerical summaries and categorical proportions, withe the first level of each category serving as the baseline.

Kaplan-Meier estimates were used to evaluate the proportion of individuals free from recurrence across different covariate groups. For individuals with no observed re-incarceration event, the time to recurrence was set to the maximum follow-up duration of 1,095 days. As illustrated in Figure 1, the survival curves for supervision type, offense type, risk ranking, and sex exhibit distinct separations, visually demonstrating the covariate-specific differences in the distribution of time to recidivism among the susceptible population. For instance, subjects with public order and violence related offenses tend to have a higher rate of recidivism.

Notably, the Kaplan-Meier curves flatten significantly toward the end of the follow-up period without converging to zero, implying an insusceptible subpopulation that standard survival models fail to capture. In addition, the observed imbalance in event occurrence suggests that the symmetric assumptions of standard link functions, like the logit, may not hold. These issues necessitate a mixture cure rate with flexible link functions that can accommodate skewed group memberships. We address these challenges by developing a Bayesian mixture cure rate framework that incorporate asymmetric link functions, as detailed in the following section.

Table 1: Summary statistics for numerical covariates.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Age	17.00	28.00	34.00	35.83	42.00	84.00
Months Supervised	0.00	4.00	7.00	12.34	13.00	536.00

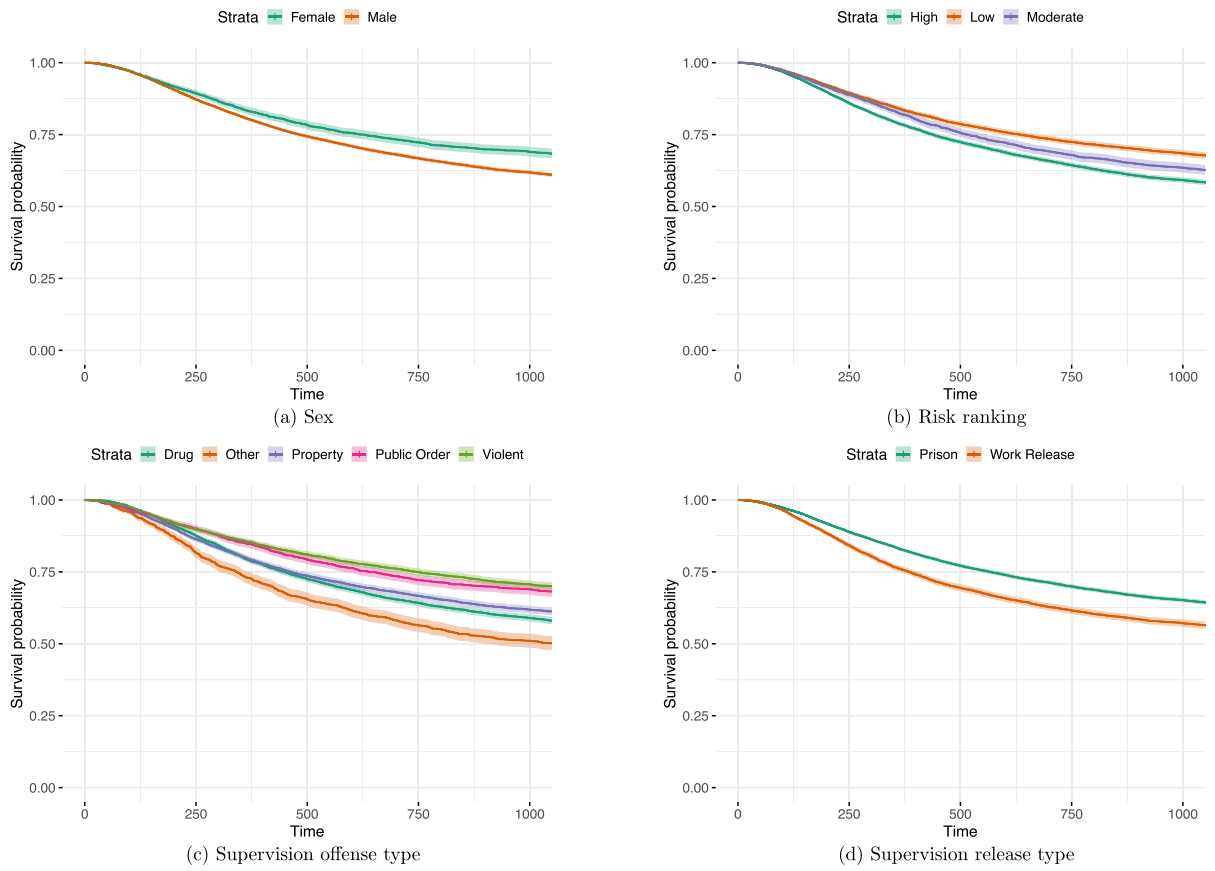
Table 2: Overview of categorical covariates.

Category	Level	Proportion (%)
Race	White	70.12
	Black	21.40
	Hispanic	5.75
	Other	2.74
Sex	Male	85.23
	Female	14.77
Supervising Unit	Fort Dodge Correctional Facility	38.63
	Clarinda Correctional Facility	10.00
	Iowa Correctional Institution for Women	14.71
	Iowa Medical Classification Center	13.91
	Mount Pleasant Correctional Facility	15.71
	Newton Correctional Facility	17.13
	North Central Correctional Facility	9.40
	Other	5.71
Supervision Type	Work Prison	70.92
	Work Release	29.08
Supervision End Reason	Parole Granted	64.05
	Discharged – Expiration of Sentence	23.79
	Released to Special Sentence	5.82
	Paroled with Immediate Discharge	2.59
	Paroled to Detainer	3.74
Offense Type	Drug	30.32
	Other	8.12
	Property	26.05
	Public Order	12.12
	Violent	23.40
Risk Ranking	High	51.54
	Low	32.55
	Moderate	15.91

3 Mixture Cure Rate Model

The mixture cure rate model is a specialized survival analysis framework designed for scenarios where a portion of the population is permanently immune from experiencing the event of interest. Unlike traditional survival models where the survival function $S(t)$ is assumed to converge to zero as $t \to \infty$, the cure model explicitly accounts for a non-zero asymptotic survival probability. The following paragraph introduces the notations used in the later sections of the paper.

For individual i, the observed time y_i is defined as the minimum of true event time T_i and



(a) Sex (b) Risk ranking

(c) Supervision offense type (d) Supervision release type

Figure 1: Kaplan-Meier curve plot by categorical variable.

censoring time C_i: $y_i = \min(T_i, C_i)$. The event status is captured by the indicator $\delta_i \in \{0, 1\}$, where $\delta_i = 1$ if the event is uncensored ($T_i \leqslant C_i$), and $\delta_i = 0$ otherwise. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$ denote the vectors of observed times and event status indicators respectively. The model introduces a latent cure indicator $w_i \in \{0, 1\}$ to classify subjects:

$$w_i = \begin{cases} 0, & \text{if subject } i \text{ is uncured,} \\ 1, & \text{if subject } i \text{ is cured,} \end{cases} \qquad i = 1, \ldots, n,$$

where $w_1, \ldots, w_n$ are assumed to be independent across subjects. The model involves independent variables for the incidence and latency components. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top \in R^{n \times p}$ represent the matrix of covariates for the incidence component with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, and $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\top \in R^{n \times q}$ represent the matrix of covariates for the latency component, with $\boldsymbol{z}_i = (z_{i1}, \ldots z_{iq})^\top$. Matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ may consist of overlapping subsets of covariates. The probability of an individual being cured is denoted by $p_{w_i} = \Pr(w_i = 1 | \boldsymbol{x}_i)$. The overall population survival and cumulative distribution functions (CDF) are $S(t)$ and $F(t)$, respectively, while $S_s(t)$ and $F_s(t)$ are the corresponding proper distribution functions for the susceptible group.

With the defined cure indicator, the decomposition for the event time T consisting of the susceptible and insusceptible individuals is (suppressing the subscript i for a generic individual)

$$T = (1 - w)T_s + w\infty,$$

where $T = \infty$ is considered that the event never happens for the cured group. The functions $F_s(t)$ and $S_s(t)$ are proper distribution functions of the event time T, meaning they are well-defined only for the susceptible group and satisfy the properties of a complete probability distribution (e.g., $S_s(0) = 1$, $\lim_{t\to\infty} S_s(t) = 0$).

We define the survival function conditional on cure status w as

$$\begin{aligned}\Pr(T \leqslant t \mid w = 0) &= F_s(t),\\ \Pr(T \leqslant t \mid w = 1) &= 0.\end{aligned}$$

The CDF for the overall population, using $\Pr(w = 1) = p_w$, becomes

$$\begin{aligned}F(t) = \Pr(T \leqslant t) &= \Pr(T \leqslant t \mid w = 1)\Pr(w = 1) + \Pr(T \leqslant t \mid w = 0)\Pr(w = 0)\\ &= (1 - p_w)F_s(t). \end{aligned} \tag{1}$$

Equivalently, the survival function for the overall population can be written as

$$S(t) = p_w + (1 - p_w)S_s(t). \tag{2}$$

This fundamental equation highlights that the overall survival probability $S(t)$ converges to the cure probability p_w as $t \to \infty$.

We specify the latency component using an AFT form,

$$\log(T_i) = \boldsymbol{z}_i^\top \boldsymbol{\beta} + \sigma\epsilon_i,$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients, σ is a scale parameter, and ϵ_i is the random error term whose distribution determines the choice of latency distribution.

4 Link Functions

The binary component of the mixture cure rate model, referred to as the incidence component, models the probability of an individual being permanently insusceptible to the event. This probability, $p_{w_i} = \Pr(w_i = 1|\boldsymbol{x}_i)$ is typically predicted using a generalized linear model framework, which applies a link function to a linear predictor based on the covariate vector $\boldsymbol{x}_i$ and the regression coefficients $\boldsymbol{\gamma}$.

The logit link is indeed one of the most common choice for this component, as it ensures that the resulting probability p_{w_i} is appropriately constrained to the interval $[0, 1]$. The probability of cure p_{w_i}, is expressed as

$$p_{w_i} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}, \quad i = 1, \ldots, n, \tag{3}$$

and applying the logit transformation yields the linear predictor

$$\text{logit}(p_{w_i}) = \log\left(\frac{p_{w_i}}{1 - p_{w_i}}\right) = \boldsymbol{x}_i'\boldsymbol{\gamma}. \tag{4}$$

While widely used, the logit link function is symmetric. The symmetry assumes an equal rate of change in the response across the range of the linear predictor. When the binary outcome is highly unbalanced, this assumption may be inaccurate. In these situations, alternative link functions based on the inverse of non-standard CDFs can better accommodate the asymmetry and nonlinearity present in the data.

4.1 Skewed and Reversed Power Logit Link

While the standard logit link function is symmetric there are instances where asymmetry in the relationship between the predictors and the response variable exists in real life. This is where the use of an asymmetric link distribution is required. The asymmetric link distributions offer greater flexibility in modeling complex relationships between predictors and binary outcomes, while effectively capturing skewness. Couple of the well-known distributions to derive the asymmetric link functions are derived from the Type I and II generalized logistic distribution (Bazán et al., 2017; Nagler, 1994).

The Type I generalized logistic distribution is also called exponentiated logistic distribution or the skewed logistic distribution. It adds an additional parameter α_s to the definition of the distribution. Simplifying the formula with $u_i = \boldsymbol{x}_i'\boldsymbol{\gamma}$, we have

$$F_{sl}(u_i) = \frac{1}{1+\exp(-u_i)^{\alpha_s}}, \qquad -\infty < u_i < \infty,\ \alpha_s > 0, \tag{5}$$

and the corresponding probability density function (pdf) is given by

$$f_{sl}(u_i) = \frac{\alpha_s \exp(-u_i)}{1+\exp(-u_i)^{\alpha_s+1}}, \qquad -\infty < u_i < \infty,\ \alpha_s > 0. \tag{6}$$

This distribution is used to generate an alternative estimator to the logit function, and is referred to as the skewed-logit, because it allows for a skewed response curve, with α serving as a parameter to measure skewness. Compared to the standard logit function, the shape of the curve changes as the value of α_s varies. For instance, for $\alpha_s < 1$, the curve becomes left skewed. The link function, denoted by slogit, is represented as

$$\text{slogit}(p_{w_i}) = \log\left\{(1-p_{w_i}^{1/\alpha_s})^{-1} - 1\right\} = \boldsymbol{x}_i'\boldsymbol{\gamma}. \tag{7}$$

Another alternative to the logit function is derived in a similar direction. Under Type II generalized logistic distribution, with an additional parameter λ_{rp}, gives the reversed exponentiated cdf form as

$$F_{rp}(u_i) = 1 - \left\{1 - \frac{1}{1+\exp(-u_i)}\right\}^{\lambda_{rp}}, \qquad -\infty < u_i < \infty,\ \lambda_{rp} > 0, \tag{8}$$

with the corresponding pdf

$$f_{rp}(u_i) = \frac{\lambda_{rp}\exp(u_i)}{\{1+\exp(u_i)\}^2}\left\{\frac{1}{1+\exp(u_i)}\right\}^{\lambda_{rp}-1}, \qquad -\infty < u_i < \infty,\ \lambda_{rp} > 0. \tag{9}$$

Finally the link function denoted by rplogit is represented by

$$\text{rplogit}(p_{w_i}) = \log\left\{(1-p_{w_i})^{-1/\lambda_{rp}} - 1\right\} = \boldsymbol{x}_i'\boldsymbol{\gamma}. \tag{10}$$

As the λ_{rp} value varies, the shape of the curve changes with some skewness. A value of $\lambda_{rp} = 1$ recovers the standard logit, while values $\lambda_{rp} < 1$ stretch the curve, creating heavier tails.

4.2 Flexible Generalized Logit Link

The flexible generalized logit (fglogit) link function is derived from the exponentiated-exponential logistic (EEL) distribution (Ghosh and Alzaatreh, 2018), which itself is based on the exponentiated-exponential distribution introduced by Gupta and Kundu (2001). This distribution combines features of both the exponentiated-exponential and logistic distributions, resulting in a flexible family capable of modeling skewness and heavy tails. With $\mu \in \mathrm{R}$ and $s > 0$ as the location and scale parameters of logistic distribution and $\alpha_{fg} > 0$ and $\lambda_{fg} > 0$ as the shape and scale parameters, the pdf and cdf are given as

$$f_{fg}(u_i) = \frac{\alpha_{fg}\lambda_{fg}\exp(u_i-\mu)/s}{s\{1+\exp(u_i-\mu)/s\}^{\lambda_{fg}+1}}\left[1-\{1+\exp(u_i-\mu)/s\}^{-\lambda_{fg}}\right]^{\alpha_{fg}-1}, \tag{11}$$

$$F_{fg}(u_i) = [1-\{1+\exp(u_i-\mu)/s\}^{-\lambda_{fg}}]^{\alpha_{fg}}. \tag{12}$$

When $\mu = 0$ and $s = 1$, specific values of λ_{fg} and α_{fg} lead to well-known distributions: when $\lambda_{fg} = \alpha_{fg} = 1$, the distribution reduces to that of the logistic distribution, while $\lambda_{fg} = 1$ results in the Type I Generalized logistic distribution, and $\alpha_{fg} = 1$ yields the Type II Generalized logistic distribution. Additionally, depending on the inequality of the parameters, the direction of skewness differs. The distribution exhibits negative skewness for $\alpha_{fg} < \lambda_{fg}$, positive skewness for $\lambda_{fg} < \alpha_{fg}$, and symmetry for $\alpha_{fg} = \lambda_{fg}$.

Utilizing the inverse of the standard EEL cdf, where $\mu = 0$, and $s = 1$, a flexible generalized logit function, denoted by fglogit, is formed (Prasetyo et al., 2020) by

$$\text{fglogit}(p_{w_i}) = \log\left\{(1-p_{w_i}^{1/\alpha_{fg}})^{-1/\lambda_{fg}}-1\right\} = \boldsymbol{x}_i'\boldsymbol{\gamma}. \tag{13}$$

For the following sections on simulation and real data application, we employ a mixture cure rate model, which combines three accelerated failure time(AFT) models: Weibull (WB), lognormal (LN), and loglogistic (LLG), for the latency component with logistic regression models using different link functions for the incidence component: Logit (L), skewed logit (S), reversed power logit (RP), and flexible generalized logit (FG). This result in 12 different models, where each model is denoted by the corresponding combination of acronyms, for example, Weibull with reversed power logit link is represented as WBRP.

5 Bayesian Inference

The proposed model combines the AFT model and logistic regression to account for both the latency and incidence components of survival data. Let $\boldsymbol{D}_{\text{obs}} = (\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n)^T = (\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{X}, \boldsymbol{Z})$ be the observed data. The incidence component is governed by the parameter vector $\boldsymbol{\nu} = (\boldsymbol{\gamma}^T, \boldsymbol{\eta}^T)^T$, where $\boldsymbol{\gamma}$ represents the regression coefficients and $\boldsymbol{\eta} = (\alpha_s, \lambda_{rp}, \alpha_{fg}, \lambda_{fg})^T$ includes additional parameters associated with the chosen link function. The latency component is governed by the parameter vector $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \boldsymbol{\sigma}^T)^T$, where $\boldsymbol{\beta}$ represents the regression coefficients and $\boldsymbol{\sigma} = (\sigma_{wb}, \sigma_{ln}, \sigma_{llg})^T$ contains scale parameters for the latency distribution. The corresponding survival function for the uncured individuals is $S_s(y_i|\boldsymbol{z}_i;\boldsymbol{\psi}) = S_s(y_i;\boldsymbol{z}_i^T\boldsymbol{\beta},\boldsymbol{\sigma})$, with density $f_s(\cdot)$. Letting $p_{w_i}(\boldsymbol{x}_i;\boldsymbol{\nu}) = \Pr(w_i = 1 \mid \boldsymbol{x}_i;\boldsymbol{\nu})$, the corresponding likelihood function with the observed data is

$$L_{\text{obs}} = \prod_{i=1}^{n}\{p_{w_i}(\boldsymbol{x}_i;\boldsymbol{\nu})f_s(y_i|\boldsymbol{z}_i;\boldsymbol{\psi})\}^{\delta_i}\{1-p_{w_i}(\boldsymbol{x}_i;\boldsymbol{\nu})+p_{w_i}(\boldsymbol{x}_i;\boldsymbol{\nu})S_s(y_i|\boldsymbol{z}_i;\boldsymbol{\psi})\}^{(1-\delta_i)}. \tag{14}$$

5.1 Model Assessment

Model performance is evaluated using three Bayesian model comparison criteria: DIC, WAIC, and LOOIC. These criteria balance model fit and complexity, providing insight into the predictive accuracy and generalizability of the models. Let $\boldsymbol{\Theta} = (\boldsymbol{\nu}^T, \boldsymbol{\psi}^T)^T = (\boldsymbol{\gamma}^T, \boldsymbol{\eta}^T, \boldsymbol{\beta}^T, \boldsymbol{\sigma}^T)^T$ denote the full vector of model parameters governing both the incidence and latency components. The observed-data likelihood in terms of $\boldsymbol{\Theta}$ is denoted as $L(\boldsymbol{\Theta} \mid \boldsymbol{D}_{\text{obs}})$.

The DIC assesses model fit by penalizing model complexity, making it suitable for comparing models with different numbers of parameters. It is computed as

$$\text{DIC} = \text{Dev}(\overline{\boldsymbol{\Theta}}) + 2p_{\text{DIC}},$$

where $\text{Dev}(\boldsymbol{\Theta}) = -2\log L(\boldsymbol{\Theta}|\boldsymbol{D}_{\text{obs}})$ is the deviance function, and $\overline{\boldsymbol{\Theta}} = \mathbb{E}[\boldsymbol{\Theta}|\boldsymbol{D}_{\text{obs}}]$ is the posterior mean for the parameters. The term $\text{Dev}(\overline{\boldsymbol{\Theta}})$ denotes the deviance evaluated at the posterior mean, while the complexity penalty $p_{\text{DIC}} = \overline{\text{Dev}(\boldsymbol{\Theta})} - \text{Dev}(\overline{\boldsymbol{\Theta}})$ represents the effective number of parameters, where $\overline{\text{Dev}(\boldsymbol{\Theta})}$ is the average deviance across posterior samples.

The WAIC is a fully Bayesian alternative to DIC, which asymptotically approximates Bayesian cross-validation. It is computed as

$$\text{WAIC} = -2\sum_{i=1}^{n} \log \mathbb{E}_{\boldsymbol{\Theta}}[p(y_i|\boldsymbol{\Theta})] + 2p_{\text{WAIC}},$$

where the first term represents the log pointwise predictive density averaged over the posterior, and $p_{\text{WAIC}} = \sum_{i=1}^{n} \text{Var}_{\boldsymbol{\Theta}}(\log p(y_i|\boldsymbol{\Theta}))$ accounts for model complexity by estimating the effective number of parameters.

The LOOIC evaluates out-of-sample predictive performance using leave-one-out cross-validation. It estimates the expected log predictive density for new observations, and is defined as

$$\text{LOOIC} = -2\sum_{i=1}^{n} \log p(y_i|\boldsymbol{y}_{-i}),$$

where $p(y_i|\boldsymbol{y}_{-i})$ represents the leave-one-out predictive density for observation i. In practice, LOOIC is approximated by Pareto-smoothed importance sampling (PSIS-LOO), which is accompanied by an estimated effective number of parameters, p_{LOOIC}, which serves as a diagnostic for model complexity (Vehtari et al., 2017). These model selection criteria are used to compare different models and determine the best fit for the given data, where smaller values indicate better model performance.

5.2 Residual Analysis

The assessment of goodness-of-fit of the models involves a Cox-Snell residual analysis (Scolas et al., 2016). The Cox-Snell residuals are given by,

$$r_{CS}(y_i) = -\log\{\widehat{S}(y_i)\} = -\log\{\hat{p}_{w_i} + (1-\hat{p}_{w_i})\widehat{S}_s(y_i|\boldsymbol{z}_i)\}, \quad i = 1, \ldots, n\ , \tag{15}$$

where $\hat{p}_{w_i} = p_{w_i}(\boldsymbol{x}_i; \hat{\nu})$ is the estimated cure probability, and $\hat{S}_s(y_i|\boldsymbol{z}_i)$ is the estimated survival function for the susceptible group. Under a good fit model, these residuals should follow an exponential distribution with rate 1, and thus, plotting the Kaplan-Meier estimate of $r_{CS}(y_i)$ against the exponential(1) reference provides a visual diagnostic for model fit.

6 Simulation Study

6.1 Simulation Setting

To evaluate the performance of our Bayesian mixture cure models under different link function settings, we conducted a simulation study assessing the 12 model configurations under two sample sizes: $n = 500$, and $n = 1000$. The configurations resulted from combining the four link functions for the cure fraction, logit (L), slogit (S), rplogit (RP), and fglogit (FG), with three parametric latency distributions: Weibull (WB), lognormal (LN), and loglogistic (LLG).

Event times for uncured individuals were simulated based on the specified latency distribution, while cure status was drawn from a Bernoulli distribution according to the calculated cure probabilities. Random uniform censoring times with Uniform(0, 100) were applied to simulate right-censoring, and individuals identified as cured were automatically assigned censored status. For each simulated dataset, covariates for the latency include an intercept and two numeric data generated from the standard normal distribution, while the incidence component include two scaled covariates generated from a uniform distribution $U(-1, 1)$, without an intercept. The true values for the parameters were set as follows: $\boldsymbol{\beta} = (1, -1, -2)^\top$, for latency regression coefficients, $\boldsymbol{\gamma} = (0.5, 0.6)^\top$, for incidence regression coefficients, $\boldsymbol{\sigma} = (1.5, 1.5, 1.5)^\top$, for latency scale for WB, LN, LLG respectively, and finally $\boldsymbol{\eta} = (1.5, 1.5, 0.5, 1.5)^\top$, corresponding to α_s, λ_{rp}, α_{fg}, λ_{fg}.

We adopted a Bayesian estimation framework, specifying independent, weakly informative priors for all model parameters to ensure regularization while maintaining flexibility. Specifically, each coefficient follows $\gamma_j \sim N(0, 1)$, $j = 1, \ldots, m$. For the link function parameters of slogit, rplogit, and fglogit, we assigned lognormal priors, such that, $\alpha_s, \lambda_{rp}, \alpha_{fg}, \lambda_{fg} \sim$ Lognormal(0, 1). These priors reflect the assumption that the shape parameters are likely near 1 while allowing flexibility in their estimation. For the latency component of the cure rate model, the scale parameter σ follows a Cauchy prior to accommodate heavy-tailed behavior, with $\sigma_{wb}, \sigma_{ln}, \sigma_{llg} \sim$ Cauchy(0, 2.5). The regression coefficient $\boldsymbol{\beta}$, which model the survival component, follow the same weakly informative normal prior as γ: $\beta_k \sim N(0, 1)$, $k = 1, \ldots, p$.

The simulation study focused on three key evaluation aspects: coverage of 95% credible intervals for parameter recovery, effectiveness of model selection criteria (WAIC, DIC, LOOIC) in identifying the true model configuration, and goodness-of-fit diagnostics. Convergence diagnostics indicated reliable estimation across all simulations, with all R-hat values close to 1 and no divergent transitions. In this paper, only the Weibull model plots are presented, while analogous results for lognormal and loglogistic latency models are included in the supplementary materials.

6.2 Scenario 1: Performance on Parameter Recovery

For the analysis of parameter recovery, a total of 1,000 replications were performed for each configuration. Figure 2 presents the overall aggregated results for the Weibull latency model across the four link functions: logit, slogit, rplogit, and fglogit. Specifically, for each parameter and link function, the plotted posterior mean estimate represents the mean of the 1,000 posterior mean estimates derived from the individual replications. Similarly, the 95% credible intervals are an aggregate interval, where the lower and upper bounds are calculated as the mean of the 2.5% and 97.5% credible interval boundaries across the 1,000 replications, respectively. Across all link functions and both sample sizes, the estimated aggregate posterior means closely align with the true values, represented by the red line, demonstrating strong point estimate accuracy.

The credible intervals are generally narrow, especially for logit, slogit, and rplogit links,

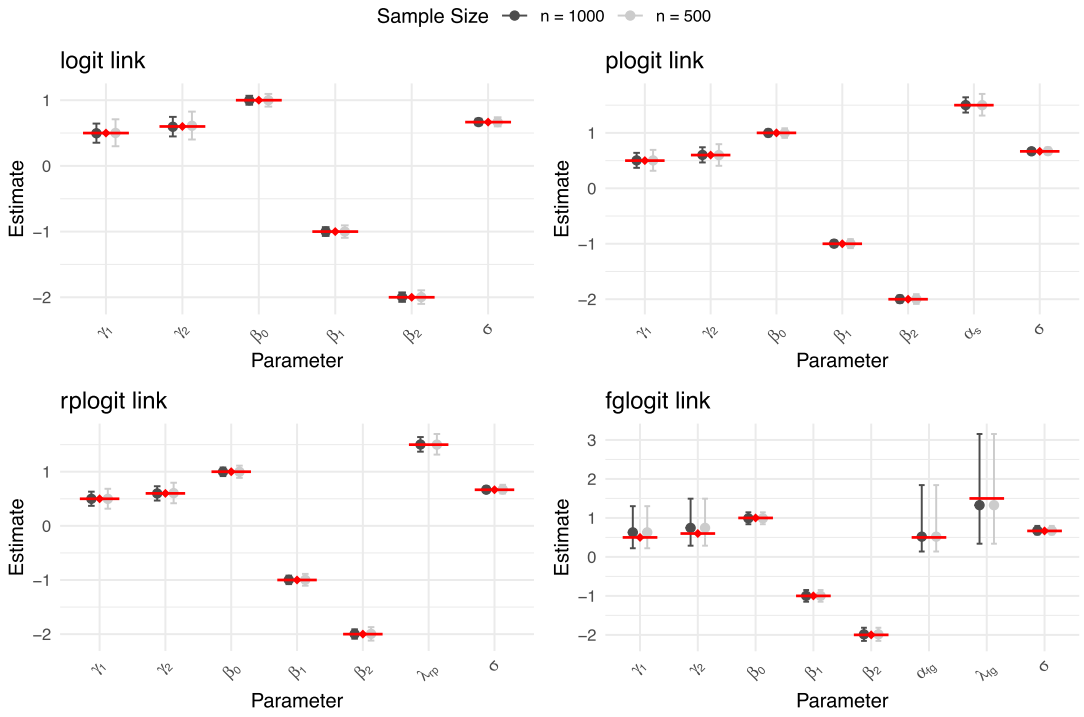


Figure 2: Posterior mean estimates and 95% credible intervals under the Weibull latency model. Black lines represent $n = 1000$, gray lines represent $n = 500$, and the red horizontal line indicates the true parameter value.

indicating high estimation precision. As expected, intervals widen under the smaller sample size ($n = 500$), while still capturing the true parameter values. The fglogit link produces wider credible intervals for incidence parameters, reflecting its greater flexibility and parameterization, which introduced additional posterior variability. This increased uncertainty arises because the model has more variability to fit the data, leading to more disperse posterior distributions for individual parameters. Despite this high variance, the model retains good recovery of the true value. While not included in this paper, similar conclusions hold under more informative priors for λ_{fg} and α_{fg}, with stable posterior estimation and acceptable interval coverage.

The coverage rates in Table 3 for all parameters are close to the nominal 95% level for both sample sizes, with slightly conservative coverage for fglogit parameters. This conservative behavior is attributed to the greater variance observed for the fglogit link in our posterior estimates, as discussed for Figure 2. Because the greater flexibility and higher parameterization of the fglogit link lead to wider credible intervals, these broader intervals are statistically more likely to include the true parameter value, resulting in the coverage probabilities that are slightly inflated above the nominal 95% level. Overall, these results confirm the accuracy and reliability of the proposed Bayesian mixture cure models under finite sample sizes.

6.3 Scenario 2: Model Assessment Validation

To evaluate the ability of model selection criteria for identifying the true model, an additional simulation study with 100 replications and $n = 1000$ was conducted. The reduced number of replications reflects the substantial computational cost of repeated posterior sampling across

Table 3: Coverage probabilities under correct model specification, where the fitted model equals the data-generating model, for Weibull latency across link functions and sample sizes.

N	Parameter	logit	slogit	rplogit	fglogit
500	γ_1	0.939	0.951	0.937	1.000
	γ_2	0.941	0.951	0.944	1.000
	β_0	0.945	0.951	0.945	0.934
	β_1	0.948	0.940	0.948	0.951
	β_2	0.944	0.953	0.947	0.961
	α	–	0.932	–	1.000
	λ	–	–	0.934	1.000
	σ	0.950	0.945	0.960	0.941
1000	γ_1	0.943	0.938	0.945	1.000
	γ_2	0.950	0.946	0.934	1.000
	β_0	0.936	0.946	0.947	0.950
	β_1	0.936	0.935	0.951	0.950
	β_2	0.934	0.939	0.940	0.934
	α	–	0.944	–	1.000
	λ	–	–	0.946	1.000
	σ	0.947	0.951	0.950	0.950

multiple model specifications (Robert and Casella, 2004). Since the objective is to assess model selection frequencies rather than estimate continuous performance measures, 100 replications provide sufficient stable estimates of selection probabilities. The data generation process remained identical to Scenario 1.

The comparison was designed to test the robustness of the model assessment criteria (LOOIC, WAIC, DIC) in choosing the correct link function over the symmetric link. For each true model configuration, the assessment values were used to select the best model out of a specific set of six competing models.

$$M_{\text{comparison}} = \{\text{True link} \times \{\text{WB, LN, LLG}\}\} \cup \{\text{Logit link} \times \{\text{WB, LN, LLG}\}\}.$$

The success rates, therefore, reflect the criteria's ability to correctly recover the true combination of both the asymmetric link function and the latency distribution over the specific misspecified alternatives. Therefore, it provides confirmation on the preference of the correct link function, even when the survival component is not perfectly specified.

Table 4 summarizes how often each criterion selected the true model out of 100 replications for the cases where the true latency was Weibull. The selection rates for Weibull-based models were exceptionally high across all criteria, often reaching 100 out of 100. However, for true lognormal and loglogistic structures, selection rates were slightly lower (full results found in the supplementary document). This decline can be attributed to the similar shape and fit of the competing distributions under the specified parameter values. Despite the decline, even in the presence of mild misspecification of the latency distribution, the model comparison criteria consistently favored the true link function over the standard logit link. These results highlight the robustness of criteria and their strong overall ability to recover the correct link function for the incidence model, even when the survival component is not perfectly specified.

Table 4: Number of times out of 100 replications true model was selected under different configurations with Weibull latency.

Criterion	slogit	rplogit	fglogit
LOOIC	100	100	99
WAIC	100	100	99
DIC	100	100	98

6.4 Scenario 3: Goodness-of-Fit Evaluation

To assess the goodness-of-fit, we generated event times from the true Weibull model with sample size of $n = 1000$, paired with each of the four link functions. We then evaluated model fit by examining the Cox-Snell residuals. For each data generating mechanism, we fitted the model using all four link functions, allowing direct comparison between the true link and alternatives.

Figure 3 presents the results when the data were generated using the symmetric logit link. In this scenario, the true model with the logit link and the proposed asymmetric link functions all provided comparable fits, as visually confirmed by the alignment of the estimated survival curves with the exponential(1) reference line. This outcome validates a key advantage of the flexible links. Since the logit link is nested within the asymmetric link functions, the flexible models can accurately reduce to the simpler logit form when symmetry is present.

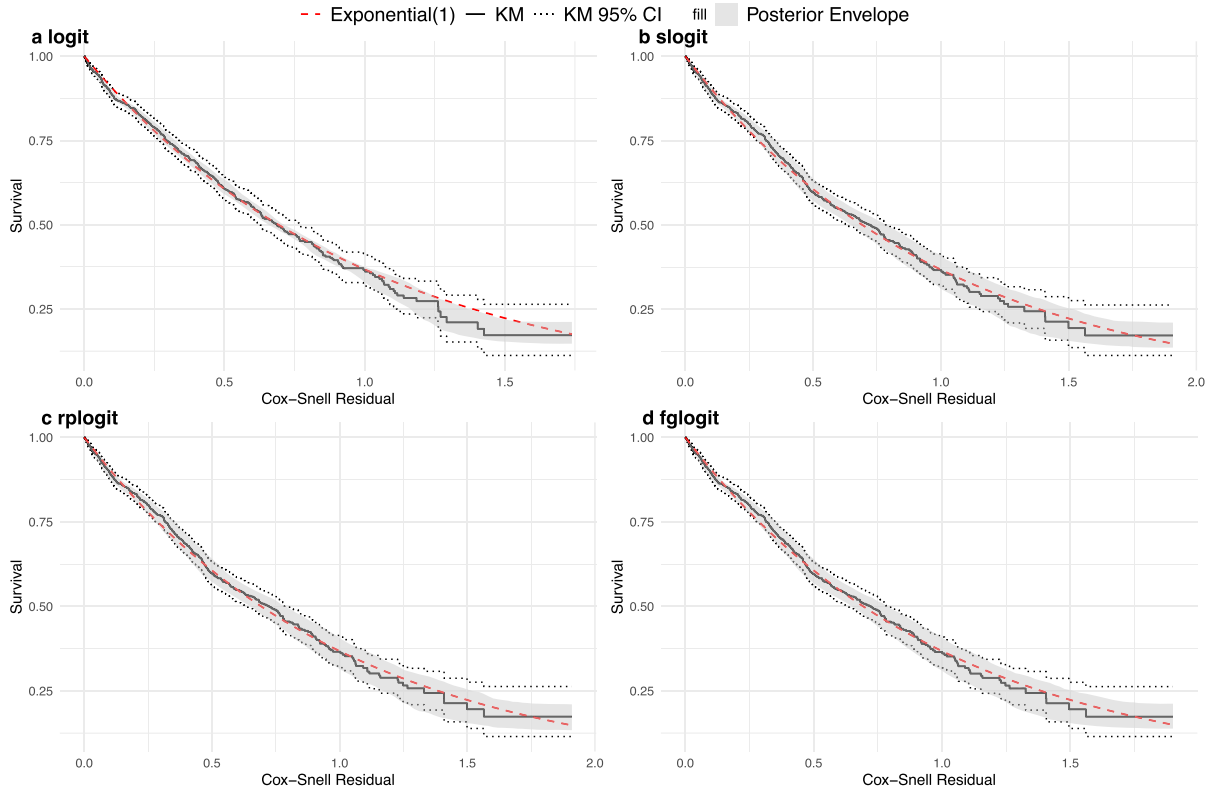


Figure 3: Cox-Snell residual plot with Weibull logit link being true model.

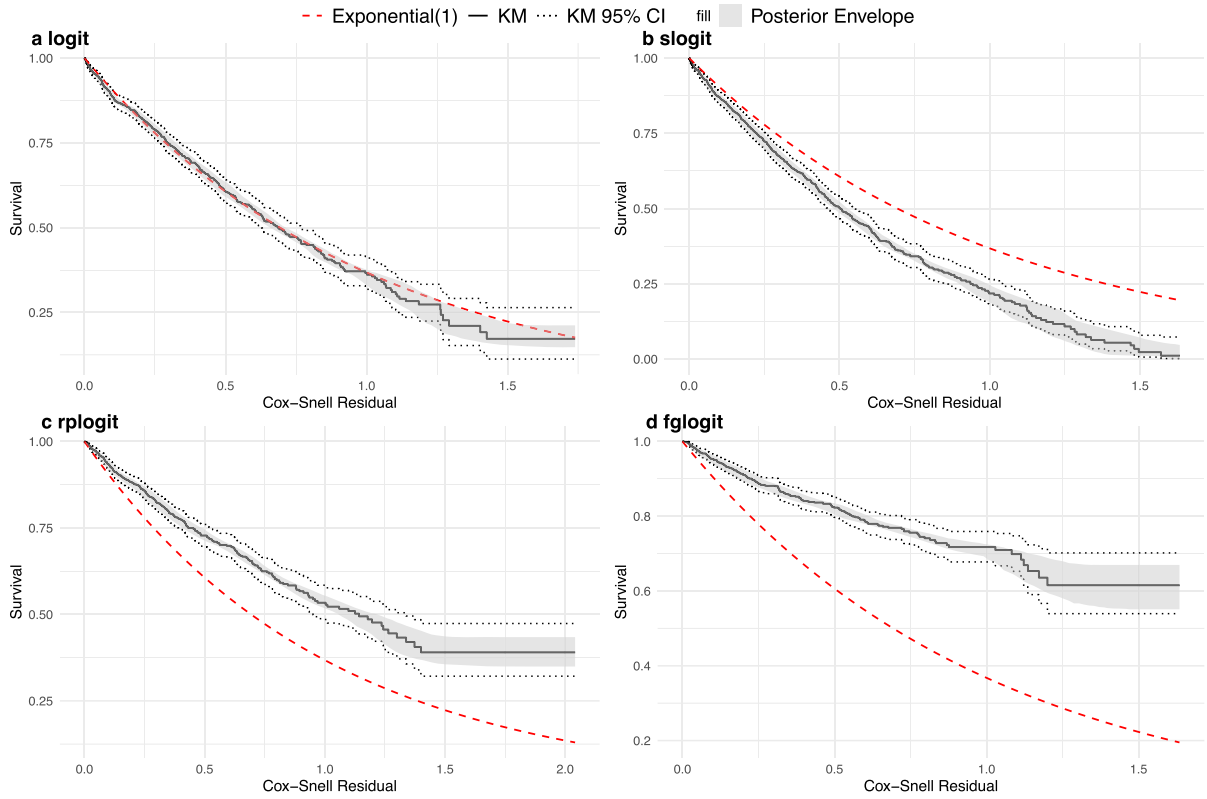


Figure 4: Cox-Snell residual plot with Weibull logit link fitted to other true link models.

In contrast, Figure 4 displays the fit when the data were generated using the asymmetric link functions but fitted with the logit link model. In these cases, the logit link model showed systematic deviation from the reference exponential(1) curve, clearly indicating lack of fit.

These results put emphasis the robustness of the proposed flexible link models, offering clear advantages when the incidence structure is skewed, and demonstrating the limitations of relying solely on the symmetric logit link when the true underlying mechanism is asymmetric. The plots presented in this paper correspond to a single simulation replicate from the Weibull case with $n = 1000$. However, similar patterns were consistently observed across other replications.

7 Application to Iowa Recidivism Data

We revisit the Iowa 3-year recidivism dataset introduced in Section 2, focusing on evaluating recidivism risk through the Bayesian mixture cure rate framework. The objective is to assess how flexible link functions perform in capturing the cure proportion and time-to-recidivism among formerly incarcerated individuals. Covariates used in the analysis include demographic and institutional variables likely to influence recidivism outcomes. The incidence model utilizes age, months under supervision, race, sex, supervising unit, supervision type, supervision end reason, and supervision offense type. The latency model used the same set of covariates, with the addition of risk ranking to model the time-to-recidivism among the susceptible population.

We apply twelve mixture cure rate models, the combinations of three AFT models (Weibull (WB), Lognormal (LN), and Loglogistic(LLG)) with four link functions (logit (L), slogit (S),

rplogit (RP), and fglogit(FG)), to predict the short-term recidivism. Using Stan, each model configuration was sampled across four independent chains with 2,000 iterations per chain, half of which were allocated to warm-up phases. The convergence of these models was evaluated through standard diagnostic checks including trace plots and assessments of the potential scale reduction factor (R-hat, Vehtari et al. (2021)). All models achieved satisfactory convergence as indicated by R-hat values close to 1.0, and there were no divergent transitions nor any instances where the maximum tree depth was saturated, indicating reliable posterior estimation.

7.1 Comparison

Table 5 compares model fit across the 12 models using LOOIC, WAIC, and DIC, where smaller values indicate better fit. The loglogistic latency model combined with the fglogit link (LLGFG) emerges as the overall best-performing model, achieving the lowest values across all three information criteria and consistently throughout the latency distributions, the fglogit link outperformed all other link functions. A crucial result is that all three asymmetric link functions consistently performed better than the symmetric logit link. This compelling evidence supports the adequacy and necessity of introducing greater flexibility into the cure fraction model. We proceed with all further inference and analysis using the LLGFG model as our final selection. Importantly, the direction and significance of the estimated regression coefficients identified in the LLGFG model were found to be highly consistent across most other model specifications, further reinforcing the robustness and reliability of the study's overall findings regarding covariate effects.

To assess goodness-of-fit of the model, we evaluated the loglogistic mixture cure rate models using posterior predictive p-values (PPP) derived based on Cox-Snell residuals from equation 15 (Gelman et al., 1996). We construct the PPP as

$$p_{\text{ppp}} = \Pr[T(r^{\text{rep}} \geqslant T(r^{\text{obs}})|\boldsymbol{y}],$$

Table 5: Bayesian assessment values with smaller values indicating better fit.

LOOIC	Weibull	Lognormal	Loglogistic
logit	32546.15	32627.34	32536.10
slogit	32544.64	32602.79	32522.44
rplogit	32544.30	32594.98	32518.52
fglogit	32538.63	32592.88	**32513.80**
WAIC	Weibull	Lognormal	Loglogistic
logit	32545.87	32627.11	32535.95
slogit	32544.32	32602.66	32522.21
rplogit	32543.99	32594.78	32518.33
fglogit	32538.32	32592.75	**32513.64**
DIC	Weibull	Lognormal	Loglogistic
logit	32494.91	32574.89	32486.40
slogit	32492.38	32550.47	32471.78
rplogit	32492.18	32544.66	32469.15
fglogit	32488.59	32542.50	**32465.76**

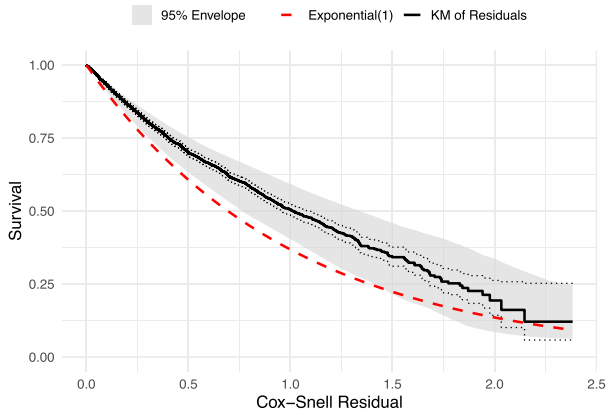


Figure 5: Kaplan Meier Cox-Snell residual plot with LLGFG.

where $T(\cdot)$ is a discrepancy measure (the integrated squared error between the empirical survival function of the residuals and the Exponential(1) survival curve), r^{rep} are the residuals computed from the posterior predictive replicates, and r^{obs} are the residuals from the observed data. A PPP close to 0.5 indicates adequate fit, while values close to 0 or 1 (under 0.05, or above 0.95) suggest model misfit. Among the loglogistic models, LLGFG model yielded a PPP of 0.558, which is close to the ideal value 0.5, suggesting an acceptable fit, indicating acceptable agreement between the model and the data. With the same loglogistic latency structure, LLGL, LLGS, and LLGRP produced PPPs ranging from 0.397 to 0.492, which are all within a reasonable range. The Cox-Snell residual plot for the LLGFG model in Figure 5 further supports the goodness-of-fit. Specifically, the empirical Kaplan-Meier curve for the residuals closely tracks the exponential reference line and remains entirely contained within the 95% posterior predictive envelope. This confirms both the overall adequacy of the model and the appropriate estimation of uncertainty. This supports the suitability of the loglogistic distribution in capturing recidivism dynamics.

7.2 Inference

Table 6 highlights the significant parameter estimates of the final LLGFG model. Similar patterns of statistical significance and effect direction were observed across all model specifications tested. The persistence of key covariates as significant predictors across varying model structures underscores the robustness of the findings and supports the reliability and interpretability of the proposed framework across a range of plausible modeling scenarios.

The event of interest in this application is recidivism, which is recommitting a crime and being reincarcerated. The time-to-event is therefore the time until the recidivism event, and cured individuals are those deemed insusceptible to the recidivism, suggesting they will never reoffend after release. The reference categories for the categorical variables are set at the first level, described in Table 2. For the latency component, a positive coefficient (β) indicates a longer expected time to the event for the susceptible population. In the incidence component, a positive coefficient (γ) indicates higher log-odds of being cured.

The latency component of the model, which describes the time-to-event for the susceptible group, yielded several significant findings. The "Month Supervised" variable is positively related to the time-to-event, with a posterior mean of 0.122. This suggests that the longer an individual is supervised, the longer it takes for them to recidivate, reflecting a protective effect of extended

Table 6: LLGFG model: posterior means and 95% HPD intervals for significant latency and cure rate variables.

Parameter	Mean	HPD 95% Interval
Latency Variables		
Month Supervised	0.122	[0.032, 0.217]
Race (ref: White)		
Black	0.187	[0.063, 0.31]
Other	−0.337	[−0.631, −0.052]
Supervising Unit (ref: Fort Dodge Correctional Facility)		
Other	−0.330	[−0.609, −0.037]
Supervision End Reason (ref: Parole Granted)		
Paroled.w.Immediate.Discharge	0.916	[0.549, 1.304]
Released.to.Special.Sentence	−0.425	[−0.629, −0.213]
Discharged.Expiration.of.Sentence	0.713	[0.553, 0.884]
Offense Type (ref: Drug)		
Property	−0.140	[−0.268, −0.01]
Risk Ranking (ref: High)		
Low	0.243	[0.122, 0.363]
Moderate	0.135	[0.001, 0.27]
σ_{llg}	1.845	[1.748, 1.946]
Incidence Variables		
Age	0.553	[0.342, 0.82]
Race (ref: White)		
Black	0.466	[0.094, 0.9]
Hispanic	0.953	[0.215, 1.731]
Supervising Unit (ref: Fort Dodge Correctional Facility)		
Clarinda.Correctional.Facility	0.569	[0.045, 1.163]
Iowa.Medical...Classification.Center	−2.767	[−3.751, −1.9]
Newton.Correctional.Facility	1.129	[0.597, 1.785]
North.Central.Correctional.Facility	0.728	[0.191, 1.339]
Other	1.212	[0.523, 2.045]
Supervision End Reason (ref: Parole Granted)		
Discharged.Expiration.of.Sentence	0.939	[0.381, 1.552]
Paroled.to.Detainer	1.997	[0.56, 3.218]
Offense Type (ref: Drug)		
Property	0.703	[0.303, 1.172]
Public.Order	1.386	[0.785, 2.117]
Violent	1.223	[0.748, 1.791]
α_{fg}	0.634	[0.436, 0.910]
λ_{fg}	0.334	[0.182, 0.568]

supervision. Significant racial differences were observed compared to the baseline White group. Black race group showed a longer time to meet the recidivism event, with posterior mean of 0.187. Compared to the reference "Parole Granted", two groups showed a significantly longer time to recidivism: "Paroled with Immediate Discharge" and "Discharged Expiration of Sentence". In contrast, those "Released to Special Sentence" had a significantly shorter time to recidivism. Also, compared to reference "Drug" offense type, subjects whose original offense was "Property" had a significantly shorter time to re-offend. Not surprisingly, lower risk classifications showed a positive effect on time-to-event. Compared to the reference high risk, those with low and moderate risk implies a longer time to recidivism.

Incidence component of the model displayed significant finding with subject's age at the original supervision start date. The positive coefficient for age means older individuals had a higher odds of being permanently non-recidivists compared to younger individuals. Consistent with the latency component, the Black race group, compared to the White group, were significantly related to higher odds of being cured, suggesting lower possibility of recidivism. This also applies to the Hispanic group. Several supervision related variables also stood out, showing differential effects on cure odds compared to the Fort Dodge Correctional Facility reference. Units such as "Clarinda", "Newton, and "North.Central" were associated with a higher odds of being cured, while the "Iowa Medical Classification Center" showed the opposite with a significant negative coefficient. Regarding the end reason for supervision, those "Discharged of Expiration of Sentence" and "Paroled to Detainer" yielded coefficients with higher odds of being immune to the event, compared to the reference "Parole Granted". Finally with offense type of "Property", "Public Order", and "Violent" compared to "Drug" all suggested higher cure odds.

The final fitted LLGFG model provides compelling statistical evidence for the use of the flexible asymmetric link function. The parameters of the fglogit link α_{fg} and λ_{fg}, have posterior means of 0.634 and 0.334 in the incidence model, respectively, with 95% credible intervals excluding 1. This statistically significant deviation from 1 implies that the fglogit link differs from all the alternative links, including the standard logit. This result suggests the structure of the Iowa recidivism 2018 cohort is better explained by an asymmetric link function, as the symmetric assumption inherent in the logit link is inadequate. Given the skewed cure fraction and imbalanced group membership observed in this application, the asymmetric link models offer a clear advantage. Overall, the flexible link models performed comparably better than the logit link, and given that the logit link is nested within these more general forms, they offer a robust and adaptable framework for incidence modeling in a wide range of applications.

8 Conclusion

This study explored Bayesian mixture cure rate models with flexible link functions to analyze recidivism outcomes in Iowa. We assessed 12 models, which combined three AFT models: Weibull, lognormal, and loglogistic, and four link functions: logit, skewed logit, reversed power logit, and flexible generalized logit link. Simulation results confirmed the robustness and flexibility of the proposed asymmetric link introduced mixture cure rate models. While the logit link performed adequately under symmetric data generating mechanisms, the flexible asymmetric links consistently offered improved fit and accuracy when the data exhibited skewness in the cure fraction, effectively recovering true parameters and demonstrating strong coverage.

In the real data application of Iowa recidivism 2018 cohort dataset, the flexible generalized logit link (LLGFG) was identified as the overall best performing model based on LOOIC, WAIC,

and DIC. The superior fit, statically supported by the finding that its link shape parameters deviated significantly from the value of 1, highlights the necessity of addressing asymmetry. The final LLGFG model revealed key covariates such as race, months supervised, offense type, and supervision conditions to be significant predictors, emphasizing the complex interplay between demographic and institutional factors on both the cure probability and the timing of recidivism.

This work validates the theoretical utility of flexible link functions and provides a robust and adaptable modeling framework for incidence modeling in survival data with a cure fraction. Future research could incorporate alternative latency structures. While parametric AFT models offer interpretability and efficiency, they may impose strict assumptions on the survival patterns. Parametric mixture distributions or semiparametric AFT models, such as those using spline-based hazard functions, could address this by allowing greater modeling flexibility. Additionally, replacing the linear predictor in the incidence component with Gaussian process priors, as suggested by Li et al. (2016), could enhance modeling of nonlinear covariate effects while maintaining flexibility. This approach could lead to more sophisticated models that better capture the complex dynamics of survival data with cure fraction.

Supplementary Material

The code supplement is available as a compressed folder. It is written in R and readily allows to replicate of the simulation results and Iowa data application. Some additional simulation results are provided within the folder.

References

Alper M, Durose MR, Markman J (2018). 2018 update on prisoner recidivism: A 9-year follow-up period, *(2005–2014). Special Report NCJ 250975*, U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, Washington, DC.

Bazán JL, Torres-Avilés F, Suzuki AK, Louzada F (2017). Power and reversal power links for binary regressions: An application for motor insurance policyholders. *Applied Stochastic Models in Business and Industry*, 33(1): 22–34. https://doi.org/10.1002/asmb.2215

Berk R (2016). *Statistical Learning from a Regression Perspective*. Springer, Philadelphia, PA, 2nd edition.

Berkson J, Gage RP (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259): 501–515. https://doi.org/10.1080/01621459.1952.10501187

Boag JW (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B, Methodological*, 11(1): 15–53. https://doi.org/10.1111/j.2517-6161.1949.tb00020.x

Chen MH, Dey DK, Shao QM (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, 94(447): 909–919. https://doi.org/10.1080/01621459.1999.10473872

De la Cruz R, Fuentes C, Padilla O (2022). A Bayesian mixture cure rate model for estimating short-term and long-term recidivism. *Entropy*, 25(1): 56. https://doi.org/10.3390/e25010056

Dressel J, Farid H (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1): eaao5580. https://doi.org/10.1126/sciadv.aao5580

Durose MR, Antenangeli L (2021). Recidivism of prisoners released in 34 states in 2012: A 5-year follow-up period, *(2012–2017). Special Report NCJ 255608*, U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, Washington, DC.

Farewell VT (1977). A model for a binary variable with time-censored observations. *Biometrika*, 64(1): 43–46. https://doi.org/10.1093/biomet/64.1.43

Farewell VT (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38(3): 1041–1046. https://doi.org/10.2307/2529885

Gelman A, Meng XL, Stern H (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4): 733–760.

Ghosh I, Alzaatreh A (2018). A new class of generalized logistic distribution. *Communications in Statistics - Theory and Methods*, 47(14): 3459–3473.

Gupta RD, Kundu D (2001). Exponentiated exponential family: An alternative to gamma and Weibull distributions. *Biometrical Journal*, 43(1): 117–130. https://doi.org/10.1002/1521-4036(200102)43:1<117::AID-BIMJ117>3.0.CO;2-R

Iowa Department of Corrections (2024). Iowa prison recidivism and change by cohort. Accessed: January 2024. Available at https://data.iowa.gov/Correctional-System/Iowa-Prison-Recidivism-and-Change-by-Cohort/dnzw-paxg/about_data.

Kim S, Chen MH, Dey DK (2008). Flexible generalized t-link models for binary response data. *Biometrika*, 95(3): 535–548. https://doi.org/10.1093/biomet/asm079

Kuk AY, Chen C (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3): 531–541. https://doi.org/10.1093/biomet/79.3.531

Li CS, Taylor JMG (2002). A semi-parametric accelerated failure time cure model. *Statistics in Medicine*, 21(21): 3235–3247. https://doi.org/10.1002/sim.1260

Li D, Wang X, Dey DK (2016). A flexible cure rate model for spatially correlated survival data based on generalized extreme value distribution and Gaussian process priors. *Biometrical Journal*, 58(5): 1178–1197. https://doi.org/10.1002/bimj.201500040

Maller RA, Zhou X (1996). *Survival Analysis with Long-Term Survivors*. Wiley.

Maltz MD (1984). *Recidivism*. Academic Press, Florida.

Nagler J (1994). Scobit: An alternative estimator to logit and probit. *American Journal of Political Science*, 38(1): 230–255. https://doi.org/10.2307/2111343

Partanen J (1969). On waiting time distributions. *Acta Sociologica*, 12(3): 132–143. https://doi.org/10.1177/000169936901200303

Prasetyo RB, Kuswanto H, Iriawan N, Ulama BSS (2020). Binomial regression models with a flexible generalized logit link function. *Symmetry*, 12(2): 221. https://doi.org/10.3390/sym12020221

Robert CP, Casella G (2004). *Monte Carlo Statistical Methods*. Springer, New York, 2nd edition.

Schmidt P, Witte AD (1989). Predicting criminal recidivism using 'split population' survival time models. *Journal of Econometrics*, 40(1): 141–159. https://doi.org/10.1016/0304-4076(89)90034-1

Schmidt P, Witte AD (2012). *Predicting Recidivism Using Survival Models*. Springer Science and Business Media.

Scolas S, Legrand C, Oulhaj A, Ghouch A (2016). Diagnostic checks in mixture cure models with interval-censoring. *Statistical Methods in Medical Research*, 27(7): 2114–2131. https://doi.org/10.1177/0962280216676502

Spiegelhalter DJ, Best NG, Carlin, BP, Van Der Linde A (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*,

64(4): 583–639. https://doi.org/10.1111/1467-9868.00353

Sy JP, Taylor JMG (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1): 227–236. https://doi.org/10.1111/j.0006-341X.2000.00227.x

Ting MH, Chu CM, Zeng G, Li D, Chng GS (2018). Predicting recidivism among youth offenders: Augmenting professional judgement with machine learning algorithms. *Journal of Social Work*, 18(6): 631–649. https://doi.org/10.1177/1468017317743137

Tollenaar N, Wartna BSJ, Van der Heijden PGM, Bogaerts S (2016). Statrec – performance, validation and preservability of a static risk prediction instrument. *BMS. Bulletin de Méthodologie Sociologique*, 129: 25–44. https://doi.org/10.1177/0759106315615504

Vehtari A, Gelman A, Gabry J (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5): 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of mcmc. *Bayesian Analysis*, 16(2): 667–718. https://doi.org/10.1214/20-BA1221

Watanabe S (2010). Asymptotic equivalence of Bayes cross-validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11: 3571–3594.

Yamaguchi K (1992). Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of 'permanent employment' in Japan. *Journal of the American Statistical Association*, 87(418): 284–292. https://doi.org/10.1080/01621459.1992.10475207

Zhang J, Peng Y (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine*, 26(16): 3157–3171. https://doi.org/10.1002/sim.2748